How a picture facilitates the process of learning from text: Evidence for scaffolding

Alexander Eitel a,*, Katharina Scheiter a, Anne Schüler a, Marcus Nyström b, Kenneth Holmqvist b

a Knowledge Media Research Center, Schleichstrasse 6, 72076 Tuebingen, Germany
b Humanities Lab of Lund University, Lund, Sweden

A R T I C L E   I N F O
Article info
Received 16 November 2012
Received in revised form 16 May 2013
Accepted 21 May 2013

Keywords:
Comprehension of text and graphics
Eye movements
Mental model construction
Picture perception
Text processing

A B S T R A C T
Three experiments were conducted to study on a more fine-grained level how processing a picture facilitates learning from text. In Experiment 1 (N = 85), results from a drawing task revealed that the global spatial structure of a pulley system picture was extracted even from its brief inspection (for 600 ms, 2 s). In Experiment 2 (N = 105), students who initially inspected the pulley system picture (for 600 ms, 2 s, or self-paced) had better comprehension of the system’s functions and made more eye movements in line with the system’s global spatial structure when listening to text than students who listened to text only. In Experiment 3 (N = 39), students who first saw the picture (for 2 s) processed written text of the pulley system’s spatial structure more efficiently than students who read text only. Results suggest that global spatial information extracted from the picture was used as a mental scaffold to facilitate mental model construction.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Many studies in recent years have shown that better comprehension can result from learning with text and pictures (i.e., multimedia) compared to learning with text only (see Anglin, Vaez, & Cunningham, 2004; Fletcher & Tobias, 2005; Levie & Lentz, 1982; Vekiri, 2002 for reviews). Beneficial effects of adding pictures to text (i.e., multimedia effects) are usually explained by comprehensive theories on multimedia learning such as the cognitive theory of multimedia learning (CTML; Mayer, 2009) or the integrative model of text and picture comprehension (ITPC; Schnotz, 2002). In the present research, multimedia effects from presenting a picture before text are examined. In particular, the interplay between text and picture processing during multimedia learning will be investigated with a focus on how information that is extracted early on when looking at a picture affects text comprehension. In the present paper we introduce a scaffolding assumption, according to which information concerning a picture’s global spatial structure is extracted even from its brief inspection. When reading subsequent text, this spatial structure is assumed to be reactivated and used as a mental scaffold to facilitate processing and comprehension of text. Three experiments are presented to test different aspects of the scaffolding assumption.

1.1. Interplay between text and picture processing in theories of multimedia learning

Both CTML (Mayer, 2009) and ITPC (Schnotz, 2002) describe multimedia learning as a multistage cognitive process during which different types of mental representation are constructed from text and pictures. These different levels of mental representation differ in the degree of elaboration. Mental representations created early during multimedia learning are linked more closely to the input modality and the concrete appearance of the representations (i.e., their surface features) than the mental representations created later in the process (i.e., an integrated mental representation or mental model).

According to CTML (Mayer, 2009), during learning with multimedia relevant information from text and pictures is first extracted and represented separately at the sensory memory level and then organized into modality-specific mental representations in working memory prior to being integrated into one coherent mental representation with the help of prior knowledge retrieved from long-term memory (Mayer, 2009). Similarly, according to the ITPC (Schnotz, 2002) when processing text the learner first constructs a
mental representation of the text surface structure, from which a propositional representation of the semantic content described in the text is generated (i.e., text base; cf. van Dijk & Kintsch, 1983).

When processing pictures, the learner first creates a perceptual representation of the visuo-spatial relations depicted in the picture that are then mapped onto semantic relations to provide the structure of the mental model (analogyic structure mapping; Schnotz & Bannert, 2003), to which verbal information can be added.

According to CTML and ITPC, interplay between text and picture processing may happen in the initial as well as in later phases during multimedia learning. With respect to the interplay at early processing stages, CTML, for instance, suggests that the modality-specific surface representations of text and pictures created at the sensory level may interact with each other. This is indicated by arrows between “words” and “sounds” in the diagrammatic representation of the theory (Mayer, 2009, p. 61). Similarly, according to the diagrammatic representation of ITPC (e.g., Schnotz, 2002, p. 109) interplay between text and picture processing is assumed at the level of constructing mental representations from text and picture: For instance, in the model arrows point from the text surface structure representation to the mental model, which is constructed from the picture, thereby indicating a potential influence from the former to the latter; similarly, an influence from the perceptual representation of the picture to the text base is represented in the model. Notably, in both CTML and ITPC these processes of early interplay between text and picture processing are hardly specified any further.

With respect to the interplay at later processing stages, CTML assumes that verbal and pictorial mental representations are integrated into one coherent mental representation with the help of prior knowledge retrieved from long-term memory. According to ITPC, there is interplay between the propositional representation and the mental model via model construction and model inspection. A mental model can be constructed from a propositional representation. Conversely, information can be read off the mental model to elaborate the propositional representation.

To summarize, according to CTML and ITPC there may be interplay between text and picture processing at different levels; however, especially the interplay that happens at early processing stages is described at a rather general level. In consequence, the present research aimed at shedding more light onto how this interplay may look like by zooming in onto one specific instance of it, namely, on how (briefly) inspecting a picture may affect processing and comprehension of subsequently read text.

How processing of text affects processing of a picture when learning with text and pictures has been investigated in a seminal study by Hegarty and Just (1993). In this study, students’ eye movements were analyzed while they were learning with text and a picture about pulley systems. The eye movement data revealed that after reading a sentence or clause, students inspected those parts in the pulley system picture about which they had just read in the text. Moreover, after reading the whole text students inspected the pulley system picture on the whole. Thus, processing of text affected processing of the picture in that the sequence of inspecting the different parts of the picture was largely guided by the text; a finding that has been confirmed in several other studies on multimedia learning (Folker, Ritter, & Sielchschmidt, 2005; Ozcelik, Arslan-Ari, & Cagiltay, 2010; Rummer, Schwepp, Fürstenberg, Scheiter, & Zindler, 2011; Schmidt-Weigand, Köhnert, & Glowalla, 2010a, 2010b). In contrast, to our knowledge less is known about how processing a picture will affect the process of learning from text.

In the following, this issue will be addressed by introducing the scaffolding assumption, which postulates that processing of a picture, even for a short time only, allows extracting information concerning its global spatial structure. When reading subsequent text, this global spatial structure will be reactivated and used as a mental scaffold to facilitate processing and comprehension of text.

1.2. The scaffolding assumption

The general notion of why there should be a mutual influence between text and picture processing is grounded in the fact that text and pictures fulfill complementary functions during multimedia learning, which together contribute to the construction of a comprehensive mental model.

In studies on learning with text and picture, participants are often required to learn about the structure and functions of causal systems such as the human circulatory system (Chi, 2000), a bicycle pump (Mayer & Moreno, 2002), or a pulley system (Hegarty & Just, 1993). Comprehension of these causal systems is usually reflected in a mental model of the system’s static spatial structure, which then allows inferring the dynamic behavior of the system’s components to better understand its functions (Hegarty, 1992; Narayanan & Hegarty, 2002). Thus, a mental model of a causal system is assumed to have a structure that is analogical to the spatial structure of its corresponding picture; however, the mental model is more abstract than the picture since it can contain abstract elements, such as information about scientific principles underlying the system’s functions, that are difficult to express in iconic representations (cf. Gyselinck & Tardieu, 1999; Johnson-Laird, 1983, 2005; Narayanan & Hegarty, 2002; Schnitz & Bannert, 2003).

Text and pictures differ on how well they support the various aspects of mental model construction for causal systems. Text is well suited to express information on a general or abstract level (Schnitz, 2002) such as information about a causal system’s functions; however, it is less suited to construct a mental model of a system’s specific spatial structure (cf. Schnitz & Bannert, 2003); moreover, the mental model of the causal system’s spatial structure constructed from text may not necessarily adequately reflect the system’s spatial structure. Since a causal system’s functions are often derived from its spatial structure (cf. Narayanan & Hegarty, 2002), constructing an inadequate mental model of a causal system’s spatial structure from text can be detrimental to comprehension of its functions (Schnitz & Bannert, 2003; Schnitz & Kürschn, 2008). On the other hand, and in line with ITPC, pictures allow a relatively direct access to mental model construction by analogical structure mapping; that is, a causal system’s specific spatial structure as expressed in the picture is mapped to the structure of the required mental model, thereby retaining its structural characteristics (cf. Glenberg & Langston, 1992; Hegarty,
These differences between text and pictures in terms of supporting mental model construction with respect to a system's spatial structure build the backbone for the scaffolding assumption.

According to the first part of the scaffolding assumption, the function of pictures is to provide direct access to a system's spatial structure already early during processing as information regarding the global spatial structure can be extracted very quickly (i.e., after a short inspection time). This is well in line with Larkin and Simon (1987), who have argued that when looking at a picture, spatial information is readily available without the need for extensive search processes. Similarly, findings from research on scene perception (Oliva, 2005) and diagram processing (Winn, 1993) suggest that briefly inspecting a causal system picture may already be sufficient to extract its spatial structure, at least on a coarse or global level. In this research, it is assumed that humans start processing a scene picture or a diagram on a coarse or global level (i.e., global precedence; Navon, 1977). Thus, the picture is initially processed as a single entity comprising the picture's global features only (Oliva, 2005; Oliva & Torralba, 2006). The longer a scene picture or diagram is processed the more perception proceeds to a local level (Antes, Penland, & Metzger, 1981; Winn, 1993), allowing for the extraction of virtual details (Liu & Jiang, 2005). This global-to-local time course of picture processing has been confirmed in a number of empirical studies (see Rayner, 2009; Winn, 1991 for reviews). Moreover, global precedence effects have been found not only with pictures of scenes (Biederman, Rabinowitz, Glass, & Stacy, 1974; Castelhano & Henderson, 2007; Fei-Fei, Iyer, Koch, & Perona, 2007; Liu & Jiang, 2005; Oliva & Schyns, 2000; Rayner, 2009; Rousselet, Joubert, & Fabre-Thorpe, 2005) but also with pictures of causal systems (Eitel, Scheiter, & Schüler, 2012). Taken together, these findings suggest that a causal system's global spatial structure can be extracted from briefly inspecting its corresponding picture.

According to the second part of the scaffolding assumption, this global spatial structure is reactivated and can be maintained active in working memory during processing of text at little cognitive cost. Representing the global spatial structure while reading is a necessary prerequisite in order to be able to explain how it can have beneficial effects on text processing and comprehension. Effects of reactivating spatial information have been shown by research conducted in the context of the blank screen paradigm (Altmann, 2004). In this research, eye movements are recorded on a blank screen while listening to text. The direction of eye movements that are made on a blank screen is assumed to reflect the spatial structure of the underlying mental representation. Using a blank screen paradigm, eye movements have been found to reflect the representation of the spatial structure constructed from both a concurrently spoken text (Johansson, Holsanova, & Holmqvist, 2006) as well as from a previously inspected picture when asked to retell the contents of the scene (Altmann, 2004; Brandt & Stark, 1997; Johansson, Holsanova, Dewhurst, & Holmqvist, 2011; Laeng & Teodorescu, 2002). Thus, when given a cue, participants' eye movements on a blank screen suggest that the spatial structure of a previously inspected picture was reactivated.

Moreover, previous research suggests that reactivated information from a picture can be maintained active in working memory to allow for the integration with information presented in subsequent text (Kulhavy, Lee, & Caterino, 1985; Kulhavy, Stock, & Kealy, 1993; McCruden, Magliano, & Schraw, 2011; Schnitz & Bannert, 2003). According to the model of working memory operations (Kulhavy et al., 1993), this integration is assumed to be successful, because a spatial representation of a coherent picture can be held as a single unit in working memory while reading subsequent text without exceeding working memory capacity. We assume that reading text about the spatial structure of a system will act as a retrieval cue that reactivates the global spatial structure already derived from the picture.

According to the third part of the scaffolding assumption, maintaining the system's global spatial structure active during reading about it will foster comprehension, which should be evident in facilitated processing of text as well as in improved learning outcomes. This is explained by assuming that briefly inspecting a causal system picture allows for a mapping of the picture's global spatial structure to the global structure of the respective mental model (cf. analogical structure mapping; Schnitz & Bannert, 2003). This means that part of the mental model (regarding the system's global structure) has already been constructed based on brief picture inspection. This part of the mental model may then serve as a mental scaffold for subsequent mental model construction from text (cf. Castelhano & Henderson, 2007; Gyselinck, Janet, & Dubois, 2008). In particular, the mental scaffold extracted from the picture constrains the range of (erroneous) interpretations that are drawn from text regarding the system's spatial structure (cf. Ainsworth, 2006; Scaife & Rogers, 1996). Namely, by providing learners with the global structure of an adequate mental model, the mental scaffold may prevent them from constructing a mental model from text that inadequately reflects the spatial structure of the system, in turn for Subsequent comprehension (Schnitz & Bannert, 2003). Accordingly, the more ambiguous a text is regarding the spatial structure of a causal system the more the mental scaffold can constrain the process of mental model construction, and thus the more helpful a brief initial picture inspection might be. Accordingly, in a study of Eitel, Scheiter and Schüler (2013), in which the presented text was ambiguous regarding the global spatial structure of the to-be-learned causal system, brief initial picture presentation fostered comprehension.

Moreover, in the study of Eitel et al. (2013) brief initial picture presentation sped up processing of text concerning the system's spatial structure, which was presented in a self-paced written manner. Since the time to process written text reflects the amount of required processing (Just & Carpenter, 1980; Kaakinen, Höynä, Keenan, 2003; Rayner, Chace, Slattery, & Ashby, 2006), this shorter processing of text was assumed to reflect facilitated processing of text about the system's spatial structure (probably in an effort to construct a mental model). According to the scaffolding assumption, facilitated processing of text about the system's spatial structure is explained by assuming that instead of having to construct a mental model from scratch when processing text only, mental model construction from text is based on the mental scaffold (i.e., global structure of the mental model) extracted from the picture. Thus, constructing a mental model about the system's global structure is not necessary anymore, leading to facilitated processing of text related to the structure of the system as reflected in shorter reading times.

1.3. Present research and hypotheses

The claims made within the scaffolding assumption were tested in three experiments. In Experiment 1 it was investigated whether information concerning the global and/or local spatial structure of a causal system picture would be extracted even after short picture inspection times, suggesting that this information is available early during processing. Following the notion of global precedence (Navon, 1977), it was assumed that processing the picture, even for a short time only, allows extracting information concerning its global spatial structure (Hypothesis 1).

In Experiment 2, it was investigated whether this global spatial structure is reactivated and maintained in working memory while processing subsequent text. In line with previous research (cf. Johansson et al., 2006; Kulhavy et al., 1993), it was assumed that
the text acts as a cue to reactivate the global spatial structure extracted from (brief) initial picture inspection. This should be reflected in students’ eye movements on a blank screen while listening to text; namely, more eye movements should be made in correspondence with the global spatial structure extracted from the picture after its (brief) inspection than when processing text only (Hypothesis 2).

In Experiments 2 and 3, it was investigated whether the reactivated global spatial structure would be helpful to comprehension. It was argued that the system’s global spatial structure extracted from (briefly) inspecting the picture fosters comprehension, because it constrains the range of (erroneous) interpretations that are drawn from text regarding the system’s spatial structure (Hypothesis 3). Thus, in Experiment 2 of the present research, in which text was ambiguous regarding the system’s spatial structure (cf. Eitel et al., 2013), processing the picture (for a short time) before text was supposed to constrain interpretation and thus to foster comprehension compared to processing just text. In Experiment 3, in which text was less ambiguous regarding the system’s global spatial structure, the effects of constraining interpretation and thus of better comprehension were assumed to be less pronounced.

In Experiment 3, regardless of text ambiguity processing of text related to the system’s spatial structure was assumed to be facilitated by a picture. Thus, as in the study of Eitel et al. (2013), participants should require less time processing text about the system’s spatial structure if a picture had been inspected for a short time before text than if no picture had been presented (Hypothesis 4).

2. Experiment 1

In Experiment 1, we addressed the first part of the scaffolding assumption by investigating whether information about a picture’s global and/or local spatial structure is extracted from its brief and self-paced inspection. Therefore, a picture of a pulley system, consisting of three pulleys that are arranged in a diagonal orientation, was presented for a short time (i.e., 600 ms and 2 s) as well as for a self-paced inspection time (see Fig. 1). Drawing scores after inspecting the picture (for a short time and for self-paced) were compared to a control condition, in which no picture was presented, to assess which information was extracted from (short) picture inspection. Due to the size of the three pulleys and their visual salience, their diagonal orientation was defined as the global spatial structure in the picture. Following Hegarty (1992), information about how the rope ran through the pulleys was defined as the local spatial structure since its extraction requires decomposition of the picture into its single components. According to Hypothesis 1, it was assumed that presenting the picture for a short time only (i.e., 600 ms and 2 s) and for self-paced inspection time allows extracting information concerning its global spatial structure (i.e., the diagonal orientation of the three pulleys).

2.1. Method

2.1.1. Participants and design

Participants were 85 students (62 female; $M_{age} = 23.67$ years, $SD_{age} = 2.96$) from a University in the South-Western part of Germany, who took part in the study for either payment or course credit. Students were randomly assigned to one of four experimental conditions (see Fig. 1): (a) control (no picture), (b) a picture presented for 600 ms (600 ms), (c) a picture presented for 2000 ms (2 s), and (d) a picture presented for as long as students liked (self-paced). There were 25 participants in the control condition, 20 participants in the 600 ms condition, 21 participants in the 2 s condition, and 19 participants in the self-paced condition.

2.1.2. Materials and procedure

In all conditions except the control condition, where students received no pictorial information, the materials consisted of a schematic picture of a pulley system in black-and-white (see Fig. 1). The picture was shown in full-size on a computer screen. Presentation of the picture was preceded by a fixation cross. Students were instructed to acquire and retain as much information as possible from the pulley system picture. In the self-paced condition, the picture remained on the screen until students pressed a key, whereas in the two remaining experimental conditions it was presented for either 600 ms or 2000 ms, respectively. After the presentation of the picture had been determined by either the subject or the system, the picture was replaced by a mask. Finally, students were required to draw the previously inspected pulley system as accurate as possible with pen on paper. In the control condition, students had to draw a pulley system based on their prior knowledge.

2.1.3. Measures and scoring

Students’ drawings were scored according to two criteria. The two criteria measured how accurately the specific spatial
structure of the depicted pulley system was drawn on a global and on a local level to indicate which information is extracted from brief and self-paced picture inspection. The first criterion referred to how well the spatial structure of the pulley system had been extracted on a coarse or global level (global accuracy; inter-rater agreement: \( r = .90, p < .001 \)). It assessed whether the orientation of the three pulleys of the pulley system picture was drawn in correspondence with the picture (see Fig. 1). In particular, one point was awarded when the upper pulley was drawn to the upper-right of the middle pulley, and another point when it was drawn to the upper-right of the lower pulley. A third point was given when the middle pulley was drawn to the upper-right of the lower pulley. There was thus a maximum score of 3 points for information about the pulley system's global spatial structure (i.e., pulley orientation).

The second scoring criterion referred to the accuracy of students' drawings at the level of single components or details (local accuracy; inter-rater agreement: \( r = .82, p < .001 \)). It was based on the scoring scheme for configuration of pulley systems from Hegarty (1992), where one point was assigned for each relation of two single components that had been correctly drawn (see Appendix A). For instance, one point was given when the upper rope ran below the middle pulley in the drawing. Another point was awarded when the upper rope was attached to the ceiling at one end in the drawing. There were a total of 11 relations that could have been drawn correctly, yielding a maximum score of 11 points for the system's local spatial structure.

### 2.2. Results

Descriptive values are reported in Table 1. Results for students' scores with regard to the pulley system's global spatial structure (i.e., orientation of pulleys) and local spatial structure were analyzed by means of two separate ANOVAs with experimental condition (control vs. 600 ms vs. 2 s vs. self-paced) as independent variable.

The ANOVA for global accuracy as dependent variable revealed a significant main effect of condition, \( F(3, 81) = 31.03, p < .001, \eta^2_p = .54 \). Repeated contrasts revealed that accuracy improved from the control to the 600 ms condition (\( p < .001 \)), but did not improve any further from the 600 ms to the 2 s (\( p = .34 \)) or from the 2 s to the self-paced condition (\( p = .23 \)). Extraction of the global spatial structure from the picture thus was as good when the picture had been inspected for 600 ms or 2 s as when picture inspection was learner-controlled.

The ANOVA for local accuracy as dependent variable revealed a significant main effect of condition, \( F(3, 81) = 41.86, p < .001, \eta^2_p = .61 \). Repeated contrasts revealed that accuracy did not improve from the control to the 600 ms condition (\( p = .93 \)) or from the 600 ms to the 2 s condition (\( p = .85 \)); rather, accuracy was higher only in the self-paced compared to the other conditions (\( p < .001 \)), demonstrating that a brief inspection of a picture was not sufficient to extract the picture's local spatial structure. Only with longer inspection of a picture, its local spatial structure could be extracted.

### 2.3. Discussion

In line with Hypothesis 1, information about the global spatial structure (i.e., diagonal pulley orientation) was extracted from brief and self-paced inspection of the pulley system picture. The local spatial structure was successfully extracted only in the self-paced condition, in which longer inspection of the pulley system picture was possible. To investigate whether the global spatial structure extracted from briefly inspecting the pulley system picture influences subsequent text comprehension, we conducted Experiment 2.

### 3. Experiment 2

Experiment 2 was conducted to test whether the pulley system's global spatial structure would be reactivated during subsequent text processing and used as a mental scaffold to foster comprehension. To ascertain that the global spatial structure was responsible for better comprehension, the pulley system picture was presented for either 600 ms or 2 s. Under these conditions, information about the global but little information about the local spatial structure had been extracted according to Experiment 1. Participants' learning outcomes in these conditions as well as their eye movements on a blank screen while listening to text were compared to those of participants who learned with text only or with text and a picture that was presented for self-paced study time before text. According to Hypotheses 2 and 3, more eye movements in line with the pulley system's global spatial structure (extracted from picture) and better comprehension of the pulley system's functions were expected from inspecting the picture for 600 ms, 2 s, and for self-paced before processing text than from processing text only.

#### 3.1. Method

#### 3.1.1. Participants and design

Data from six participants had to be excluded from the analysis because of problems with eye tracking (e.g., low validation accuracy). Data from 105 students (74 female; \( M_{\text{age}} = 23.93 \) years, \( SD_{\text{age}} = 3.60 \)) from a University in the South-Western part of Germany were analyzed. Students were either paid 8 Euros or received course credit as compensation for participating. They were randomly assigned to one of five experimental conditions (see Fig. 2): (a) text-only, (b) a picture presented for 600 ms before the text (600-before), (c) a picture presented for 2000 ms before the text (2 s-before), (d) a picture presented for as long as they liked before the text (self-paced before), or (e) text and picture presented simultaneously. Although learning outcomes were assessed in all conditions, the latter condition was introduced only to determine the pattern of eye movements that corresponds to the pulley system's global spatial structure while listening to the respective text (for details see ‘Measures’ section below). There were 20 participants in the text-only condition, 21 participants in the 600-before condition, 23 participants in the 2 s-before condition, 20 participants in the self-paced before condition, and 21 participants in the simultaneous condition.

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Control</th>
<th>600 ms</th>
<th>2 s</th>
<th>Self-paced</th>
<th>M (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global spatial structure (min. – 0, max. – 3)</td>
<td>0.32 (0.56)</td>
<td>2.10 (1.20)</td>
<td>2.38 (1.16)</td>
<td>2.74 (0.65)</td>
<td>1.79 (0.34)</td>
</tr>
<tr>
<td>Local spatial structure (min. – 0, max. – 11)</td>
<td>0.84 (1.63)</td>
<td>0.90 (0.79)</td>
<td>0.76 (0.70)</td>
<td>7.47 (4.31)</td>
<td>2.32 (3.57)</td>
</tr>
<tr>
<td>Picture inspection time (in sec)</td>
<td>0 (0)</td>
<td>0.60 (0)</td>
<td>2.00 (0)</td>
<td>28.64 (15.97)</td>
<td>7.04 (13.83)</td>
</tr>
</tbody>
</table>
3.1.2. Materials

Participants in all conditions received a spoken verbal description of the pulley system’s structure and functions; moreover, in four out of five conditions this spoken text was either preceded or accompanied by the pulley system (see Fig. 2). The text was a slightly extended, German version of the text (240 words) used by Hegarty and Just (1993) about the structure and functions of a pulley system (see Appendix B). In the first part of the text (called structure section thereafter) the spatial structure of the pulley system used in the present experiment was described (e.g., “the upper pulley is attached to the ceiling”). As in its original version (Hegarty & Just, 1993), the text contained information about all of the pulley system picture’s components and their relations; however, it did not contain any information about the diagonal orientation of the three pulleys of the system, namely that the upper pulley is located to the upper-right of the middle pulley, and the middle pulley to the upper-right of the lower pulley. Rather, the text described the upper pulley merely to be located above the middle pulley, and the middle pulley to be located above the lower pulley. Information about the diagonal orientation of the three pulleys thus could be extracted only from the picture.

In the second part of the text (called function section thereafter) it was described what happens when the rope is pulled (e.g., “...middle pulley lifted”; cf. Boucheix & Schneider, 2009; Hegarty & Just, 1993) and how the underlying principle of pulley systems in general works (i.e., each free pulley splits the force with which the weight has to be lifted in half and doubles the length of rope to be pulled). The explanation about the principle underlying pulley systems was given in the text and could not be inferred from the picture.

3.1.3. Measures and scoring

A shortened paper-pencil version of the paper folding test (PFT; Ekstrom, French, & Harman, 1976) was administered to test for students’ spatial abilities, since spatial abilities have been shown to influence students’ eye movements (Kozhevnikov, Motes, & Hegarty, 2007) as well as their ability to mentally animate causal diagrams (Hegarty, 1992, 2004). To assess students’ conceptions of pulley systems, they were given a blank sheet of paper on which they were asked to draw a pulley system, both before (premodels; Butcher, 2006) and after being presented with both picture and text about the pulley system (postmodels). Drawings scores prior to the learning phase were assessed to control for whether the spatial structure of students’ prior conceptions about pulley systems matched the spatial structure of the to-be-learned pulley system in the present experiments. Both premodels and postmodels were scored in a similar manner than in Experiment 1, namely with respect to the degree to which the system’s global and local spatial structure had been extracted (see Appendix A). With regard to the system’s global spatial structure, premodels and postmodels were scored according to two criteria. First, as in Experiment 1, it was assessed whether the orientation of the three pulleys of the pulley system picture was drawn in correspondence with the picture (i.e., from bottom-left to top-right). Second, it was assessed whether the orientation of the three pulleys was drawn in the opposite direction (i.e., from bottom-right to top-left). For both scoring schemes, and for both premodels and postmodels, with regard to the global spatial structure a minimum of 0 points and a maximum score of 3 points could be obtained. With regard to the local spatial structure, premodels and postmodels were scored in the same manner as in Experiment 1 so that a minimum score of 0 points and a maximum score of 11 points could be achieved. Inter-rater agreement on 25% of the data was sufficiently high with regard to both global ($r = .85, p < .001$) and local spatial information ($r = .86, p < .001$) so that only one rater continued scoring the data.

Verbal recall of the system’s global spatial structure (pulley orientation) was assessed with 6 verification items (e.g., “the upper pulley is located more to the left than the middle pulley”; no was correct), and verbal recall of the local spatial structure with 11 items (e.g., “both ropes are attached to the ceiling with one end”; yes was correct). Each correct response in the verbal verification test was credited with 1 point. After each response, students had to mark how confident they felt with regard to their response on a five-point Likert scale with the following options: “guessed”, “uncertain”, “a little uncertain”, “certain”, “very certain”. If students marked the lowest score in their confidence rating (“guessed”), the respective response was multiplied by 0 so that guessed responses were not counted in the analysis. All responses with a confidence rating that was higher than “guessed” were counted in the analysis (multiplied by 1). This was done to bypass the problem of guessing probability leading to an increased reliability of the test (cf. Cierniak, Schetler, & Gerjets, 2008; Conway, Gardiner, Perfect, & Cohen, 1997). The maximum score was 6 points for verbal recall of the global spatial structure, and 11 points for verbal recall of the local spatial structure.

Fig. 2. Design of Experiment 2. Each column represents an experimental condition.
Comprehension was assessed by means of a mental animation test that required knowledge about relations between single components in the system (i.e., local spatial information) and questions regarding the principle underlying pulley systems, namely that each free pulley splits the weight to be lifted in half and doubles the length of rope to be pulled (i.e., principle of free pulleys). The mental animation test comprised 9 verification items (e.g., “if the free end of the upper rope is let go, then the middle pulley turns clockwise”; yes was correct), so that students could score a maximum of 9 points. Comprehension of the principle of free pulleys was assessed via 7 verification items (e.g., “if the weight was attached to the middle pulley, then the rope would have to be pulled with the same force than when the weight is attached to the lower pulley”; no was correct) and via 4 labeling items. For these items, students had to label how much the to-be-lifted weight is reduced in a depiction of a specified pulley system compared to another one. Depictions of pulley systems differed in their number and spatial arrangement of free pulleys so that correctly solving the task required application of the principle of free pulleys to different depictions of pulley systems. Thus, conceptual knowledge retained from the instruction had to be used to solve a new problem (cf. transfer; Mayer, 2002). Results for both the verification test and the labeling test were merged so that students could score a maximum of 11 points for comprehension of the principle.2

Participants’ eye movements were recorded while they were listening to spoken text and looking at a blank screen (in conditions with picture-before-text and in the text-only condition) or at the picture of the pulley system (in the simultaneous condition), respectively. According to Hypothesis 2, more eye movements in line with the pulley system’s global spatial structure (i.e., diagonal pulley orientation; only presented in the picture) were expected in conditions with brief and self-paced picture presentation before text compared to the text-only condition. To determine whether eye movements were in line with the pulley orientation in the picture, we first analyzed the direction of saccades on the pulley system picture in the simultaneous condition (see Fig. 3). In this condition, most saccades had an orientation of 45°–75°, and very few saccades were made in the opposite diagonal direction (i.e., had an orientation of 105°–135°; see Fig. 4). Accordingly, the pattern of performing more saccades in the direction of 45°–75° (corresponding saccades) than saccades in the direction of 105°–135° (non-corresponding saccades) was defined as being in accordance with the pulley orientation as depicted in the picture. To be able to statistically compare whether more saccades were made in the corresponding than in the non-corresponding direction in the different experimental conditions, we first computed the standardized relative frequency3 of corresponding and non-corresponding saccades, and then subtracted the frequency values for non-corresponding saccades from the values for corresponding saccades for each individual. As a result, we obtained one measure for whether students’ mental models of the system’s spatial structure matched the global spatial orientation of the picture.

3.1.4. Procedure

Students were tested in single sessions of approximately 50 min. Students first had to draw a picture of a pulley system and fill in a demographic questionnaire in paper pencil format. Then, in all except for the text-only condition the pulley system picture was presented on a computer screen, preceded by a central fixation cross. In the simultaneous condition, spoken text was presented concurrently with the picture for 105 s. In

2 Note that care was taken that all items for verbal recall and postmodel of the system’s local spatial structure as well as all items for mental animation and comprehension of the principle could be answered correctly even without information about the diagonal orientation of the three pulleys that could be extracted only from the picture in Experiment 2. Thus, possible effects do not merely reflect a manipulation check.

3 We computed the standardized relative frequency of saccades to reduce un-systematic variance stemming from large between-participants differences with regard to the total number of saccades performed on the blank screen.
conditions with picture before text (see Fig. 2), the picture appeared on the screen for 600 ms or 2 s or the picture remained visible on the screen until students signaled that they had sufficiently inspected the picture (self-paced). The experimenter responded to the signal by pressing a key. Afterwards, the picture was replaced by a visual mask. The visual mask was presented to ascertain that subsequent eye movements on the blank screen are not driven by a lingering sensory representation of the picture (e.g., an after-image). After the mask had disappeared, students saw a blank white screen, and heard the spoken text about the pulley system.

In all conditions, students received an instruction about how material would be presented to them. In conditions in which text was presented with a blank screen, students were told that the objective of the eye movement recording on the blank space was to measure the pupil dilation (cf. Johansson et al., 2006). After learning from text and picture about pulley systems, students in all conditions were instructed to draw a picture. Subsequently, students in all conditions completed the verbal verification tests, followed by the labeling tests and the PFT.

3.1.5. Apparatus
Stimuli were presented on a monitor that had a 1280 × 1024 pixel-resolution, a size of 29.54 × 24.37 degrees of visual angle, and a refresh rate of 60 Hz. Eye movements were recorded by means of a video-based eye tracker (High-speed 1250; iViewX 2.4™) from SensoMotoric Instruments (SMI) with a 500 Hz sampling rate. The system was calibrated and validated using a 13-point calibration image. Fixations and saccades were detected with the saccade velocity detection algorithm from the default settings of BeGaze3.0™. Stimuli were presented using E-prime 2.0 Professional from Psychology Software Tools.

3.2. Results
We first assessed whether participants were comparable with regard to their spatial ability and premodels, followed by analyses of participants’ learning outcomes and eye movements across experimental conditions (text-only, 600-before, 2 s-before, self-pace).

Note: There are no saccades with an orientation of 180° to 360° in the graphs, because they had the same orientation as saccades in the exact opposite direction. For instance, a saccade made in the orientation 240° was the same as a saccade in the orientation of 60°, only that the saccade was made either from top-to-bottom or from bottom-to-top. Since we were not interested in this difference, we merged saccades with corresponding orientations.

Fig. 4. Rose diagrams of the standardized relative frequencies of students’ saccades made in the various directions while listening to text about the structure and functions of the pulley system in the five experimental conditions in Experiment 2.

4 Data was analyzed only from participants whose validation accuracy was better or equal to a mean deviation of 1 degree of visual angle in both the vertical and horizontal direction.
cased before). Subsequent planned comparisons were conducted to test Hypotheses 2 and 3. Accordingly, planned comparisons were expected to differ significantly from the text-only condition to the conditions with initial picture presentation for 600 ms, 2 s, and self-paced.

3.2.1. Premodels and spatial abilities

Descriptive values are shown in Table 2. With regard to participants’ spatial abilities (i.e., PFT scores), a one-way ANOVA with experimental condition as between-subjects factor (text-only vs. 600-before vs. 2 s-before vs. self-paced before) revealed no main effect of condition, \( F < 1 \). With regard to participants’ premodes (drawings prior to learning phase), three separate one-way ANOVAs with experimental condition as between-subjects factor (text-only vs. 600-before vs. 2 s-before vs. self-paced before) did not reveal a main effect of condition for the global spatial structure in the corresponding direction, \( F(3, 80) = 1.30, p = .28, \eta^2_p = .05 \), for the global spatial structure in the non-corresponding direction, \( F(3, 80) = 1.34, p = .27, \eta^2_p = .05 \), or for the local spatial structure, \( F < 1 \). Participants thus did not differ between experimental conditions with regard to their spatial abilities or premodes. With regard to the premodes of the global spatial structure, most participants did not represent the pulley system in a particular diagonal direction (74 out of 84 participants scored 0). Moreover, the low scores for premodes for the system’s local spatial structure suggest that students’ prior conceptions of pulley systems were not very elaborate.

3.2.2. Learning outcomes

We analyzed learning outcomes separately for each of the dependent variables; these are postmodels, verbal recall, mental animation, and comprehension of the principle underlying pulley systems. Descriptive values are reported in Table 2.

**Postmodels.** Analyzing scores for students’ postmodels by means of an ANOVA revealed significant main effects of experimental condition both with regard to the global spatial structure in the corresponding direction, \( F(3, 80) = 21.93, p < .001, \eta^2_p = .45 \), the global spatial structure in the non-corresponding direction, \( F(3, 80) = 3.54, p = .02, \eta^2_p = .12 \), and with regard to the local spatial structure, \( F(3, 80) = 16.83, p < .001, \eta^2_p = .39 \). Planned comparisons further revealed that postmodels for the global spatial structure in the corresponding direction were more accurate when the picture was presented for 600 ms, 2 s and for self-paced before text than in the text-only condition (all \( p < .05 \)). For the global spatial structure in the non-corresponding direction, postmodel scores were higher in the 600-before than in the text-only condition. There were no further significant differences. For the local spatial structure, postmodels were more accurate in the condition with self-paced picture presentation before text than in the text-only condition (\( p < .001 \)); there were no significant differences between conditions with 600 ms- and 2 s-picture presentation and the text-only condition (both \( p > .05 \)).

**Verbal recall.** The analysis of scores for verbal recall across experimental conditions revealed significant main effects both for global, \( F(3, 80) = 13.78, p < .001, \eta^2_p = .34 \), and for local spatial information, \( F(3, 80) = 6.42, p = .001, \eta^2_p = .19 \). In line with results for postmodels, planned comparisons revealed that in all three conditions where the picture had been presented prior to text (for 600 ms, 2 s, and self-paced) global spatial information was extracted to a higher degree than when only text had been presented (all \( p < .05 \)). With regard to local spatial information, planned comparisons revealed that presenting the picture self-paced before text led to better recall than presenting text only (\( p = .01 \)), whereas presenting the picture for 600 ms and 2 s-before text did not (both \( p > .05 \)).

**Mental animation.** Concerning mental animation performance, the ANOVA revealed a significant main effect of experimental condition, \( F(3, 80) = 4.22, p = .008, \eta^2_p = .14 \). Planned comparisons further revealed that only with self-paced presentation of the picture before text, mental animation was better than with text only (\( p = .002 \)). Presenting the picture for either 600 ms or 2 s before text did not lead to better mental animation performance than presenting text only (both \( p > .05 \)).

**Comprehension of principle.** The ANOVA with comprehension of the principle as dependent variable revealed a significant main effect of experimental condition, \( F(3, 80) = 2.86, p = .04, \eta^2_p = .10 \). Planned comparisons revealed that in all three conditions where the picture had been presented prior to text (for 600 ms, 2 s, and self-paced) comprehension of the principle was better than in the text-only condition (all \( p < .05 \)).

3.2.3. Eye movements

Descriptive values are shown in Table 3. We analyzed whether students’ saccades were made in correspondence with system’s global spatial orientation as depicted in the picture while they were listening to text and looking at a blank screen. Therefore, we included the difference scores for the relative standardized frequency of corresponding saccades (i.e., according to pulley orientation from picture: 45°–75°) minus non-corresponding saccades (i.e., 105°–135°) as a dependent variable in an ANOVA with experimental condition (text-only vs. 600-before vs. 2 s-before vs. self-paced before) as between-subjects factor. Results revealed a significant main effect of condition, \( F(3, 80) = 2.94, p = .04, \eta^2_p = .10 \). Subsequent planned comparisons revealed that values of the difference scores were significantly higher in all three conditions with picture presentation before text (for 600 ms, 2 s, or self-paced) than in the text-only condition (all \( p < .05 \)). Thus, the pattern of eye movements made while listening to text were more in line with the global spatial orientation of the pulley system picture in conditions with brief and self-paced picture inspection than with text presentation only. This suggests that constructing a mental representation from text was influenced by the system’s global spatial structure, which had been extracted from previous picture inspection and reactivated during text comprehension. Moreover, results revealed that students in the text-only condition made as many saccades in the corresponding as in the non-corresponding direction, \( \mu(19) = -0.11, p = .92 \). If part of the students in the text-only condition had constructed a mental model in the non-corresponding direction and the other part of the students in the corresponding direction, then one would expect a bimodal distribution, in which case there should be no normal distribution. Thus, we tested whether mean diagonal orientation of saccades was normally distributed across participants. Results showed that the data were not normally distributed (Kolmogorov-Smirnov-Test, \( ks(20) = -0.11, p < .20 \)), suggesting that most students in the text-only condition did not represent any of the two diagonal orientations. Descriptive values are shown in Table 4.

3.3. Discussion

Results for postmodels and verbal recall replicate findings from Experiment 1 in that the system’s global spatial structure was...
Table 2
Means (standard deviations) for spatial abilities, premodels, and learning outcomes as a function of experimental condition in Experiment 2.

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Required information</th>
<th>Text-only</th>
<th>600-Before</th>
<th>2 s-Before</th>
<th>Self-paced before</th>
<th>Simultaneous</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial abilities (min. = 0, max. = 10)</td>
<td>–</td>
<td>6.75 (2.34)</td>
<td>7.00 (1.92)</td>
<td>6.61 (2.19)</td>
<td>6.85 (1.87)</td>
<td>6.36 (2.50)</td>
<td>6.71 (2.15)</td>
</tr>
<tr>
<td>Preamodel (min. = 0, max. = 3)</td>
<td>Corresponding</td>
<td>0.16 (0.37)</td>
<td>0.48 (0.22)</td>
<td>0.35 (0.78)</td>
<td>0.16 (0.49)</td>
<td>0.18 (0.51)</td>
<td>0.16 (0.37)</td>
</tr>
<tr>
<td>Preamodel (min. = 0, max. = 3)</td>
<td>Non-corresp.</td>
<td>0.05 (0.22)</td>
<td>0.14 (0.36)</td>
<td>0.04 (0.21)</td>
<td>0.00 (0.00)</td>
<td>0.48 (0.22)</td>
<td>0.06 (0.23)</td>
</tr>
<tr>
<td>Postmodel (min. = 0, max. = 3)</td>
<td>Corresponding</td>
<td>0.25 (0.44)</td>
<td>2.00 (1.34)</td>
<td>1.61 (1.31)</td>
<td>2.85 (0.67)</td>
<td>2.81 (0.68)</td>
<td>1.91 (1.33)</td>
</tr>
<tr>
<td>Postmodel (min. = 0, max. = 3)</td>
<td>Non-corresp.</td>
<td>0.05 (0.22)</td>
<td>0.62 (1.20)</td>
<td>0.00 (0.00)</td>
<td>0.15 (0.67)</td>
<td>0.00 (0.00)</td>
<td>0.16 (0.65)</td>
</tr>
<tr>
<td>Verbal recall (min. = 0, max. = 6)</td>
<td>Corresponding</td>
<td>2.50 (1.15)</td>
<td>4.00 (1.30)</td>
<td>3.83 (1.56)</td>
<td>5.05 (0.89)</td>
<td>5.57 (0.51)</td>
<td>3.85 (1.53)</td>
</tr>
<tr>
<td>Verbal recall (min. = 0, max. = 6)</td>
<td>Non-corresp.</td>
<td>0.89 (1.22)</td>
<td>1.38 (1.96)</td>
<td>0.87 (1.42)</td>
<td>1.28 (0.99)</td>
<td>0.38 (0.74)</td>
<td>0.66 (1.34)</td>
</tr>
<tr>
<td>Mental animation (min. = 0, max. = 10)</td>
<td>Global spatial structure</td>
<td>4.10 (1.74)</td>
<td>4.76 (1.48)</td>
<td>4.22 (2.35)</td>
<td>5.95 (1.67)</td>
<td>5.86 (1.68)</td>
<td>4.47 (0.96)</td>
</tr>
<tr>
<td>Prehear (min. = 0, max. = 10)</td>
<td>Local spatial structure</td>
<td>3.65 (1.81)</td>
<td>4.91 (1.55)</td>
<td>5.26 (2.24)</td>
<td>4.90 (1.89)</td>
<td>6.91 (1.55)</td>
<td>4.70 (1.96)</td>
</tr>
</tbody>
</table>

Note: Better scores on these tests with (brief initial) picture inspection than with text only reflect a manipulation check, since the information required to correctly respond to the test items could be extracted only from the picture, and not from text.

Table 3
Means (standard deviations) for the standardized relative frequency of saccades made according to the pulley orientation in the picture (corresponding) minus saccades made in the opposite direction (non-corresponding) as a function of experimental condition in Experiment 2.

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Required information</th>
<th>Text-only</th>
<th>600-Before</th>
<th>2 s-Before</th>
<th>Self-paced before</th>
<th>Simultaneous</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saccades in corresponding – non-corresponding direction</td>
<td>Corresponding</td>
<td>–0.00 (.10)</td>
<td>–0.08 (.15)</td>
<td>–0.07 (.11)</td>
<td>–0.10 (.11)</td>
<td>–0.32 (.10)</td>
<td>–0.11 (.16)</td>
</tr>
</tbody>
</table>

4. Experiment 3

Experiment 3 was conducted to investigate Hypotheses 3 and 4, namely that the pulley system's global spatial structure extracted from brief initial picture inspection facilitates comprehension and processing of text. In particular, in Experiment 3 it was tested whether the beneficial effects on comprehension from brief initial picture inspection (i.e., for 2 s) would hold up even if text is made more specific regarding the system's global spatial structure (i.e., by including 8 direction words). According to Hypothesis 3, beneficial effects on comprehension from brief initial picture presentation were expected to be less pronounced than in Experiment 2. To investigate facilitated processing of text, text was presented in a written rather than in a spoken format to allow for self-paced reading. As in the study of Eitel et al. (2013), facilitated processing of text was operationalized via shorter processing of text. According to Hypothesis 4, participants were expected to require less time processing text about the system's spatial structure if they had inspected the picture for a short time before text (i.e., for 2 s) compared to when they learned with text only.

4.1. Method

4.1.1. Participants and design

Data from three students had to be excluded from the analysis because of problems with eye tracking. Data from 39 students (26 female; Mage = 23.21 years, SDage = 3.75) from a University in the South-Western part of Germany were analyzed. Students were either paid 8 Euros or received course credit as compensation for participation. Students were randomly assigned to one of two experimental conditions: (a) text-only or (b) a picture presented for 2000 ms before the text (2 s-before). For economic reasons we did not implement a 600 ms condition any longer. There were 20 students in the text-only condition, and 19 students in the 2 s-before condition.

Table 4
Number of participants as a function of their mean saccade orientation while processing text in the respective experimental conditions in Experiment 2.

<table>
<thead>
<tr>
<th>Experimental condition</th>
<th>Mean saccade orientation smaller than –15°</th>
<th>Mean saccade orientation between –14° and –5°</th>
<th>Mean saccade orientation between –4.9° and 5°</th>
<th>Mean saccade orientation between 5.1° and 15°</th>
<th>Mean saccade orientation larger than 15°</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text-only</td>
<td>1</td>
<td>5</td>
<td>9</td>
<td>5</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>600-Before</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>9</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>2 s-Before</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>23</td>
</tr>
<tr>
<td>Self-paced before</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>20</td>
</tr>
<tr>
<td>Simultaneous</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>19</td>
<td>21</td>
</tr>
</tbody>
</table>
4.1.2. Materials

Materials were the same as in Experiment 2 with two exceptions. Written rather than spoken text about the structure and functions of pulley systems that fitted on a single page on the computer screen was presented to students. Compared to the text in Experiment 2, eight words about the diagonal orientation of pulleys were added to the text (248 words in total). In particular, the text version used in Experiment 3 described that the upper pulley is located to the upper-right of the middle pulley etc., so that information about the diagonal orientation of pulleys could be extracted both from picture and from text.

4.1.3. Measures and scoring

Measures and scoring for premodels, spatial abilities and learning outcomes were the same as in Experiment 2. Inter-rater agreement on scores for postmodels were sufficiently high for both global spatial information ($r = 1.00$, $p < .001$) and local spatial information ($r = .83$, $p < .001$). Students’ processing of written text about the structure and functions of the pulley system was analyzed by assessing the dwell times (summed duration of all fixations and saccades) regarding the structure and the function section in the text.

4.1.4. Apparatus

Apparatus was identical to Experiment 2.

4.1.5. Procedure

The procedure was the same as in Experiment 2, except that written rather than spoken text was presented in Experiment 3.

4.2. Results

We first assessed whether participants were comparable with regard to their spatial ability and premodels across experimental conditions (text-only vs. 2 s-before), followed by analyses of participants’ learning outcomes and eye movements. Descriptive values are reported in Table 5.

4.2.1. Premodels and spatial abilities

Students in the text-only condition were comparable to students in the 2 s-before with regard to their spatial abilities (i.e., PFT scores), $F < 1$, as a one-way ANOVA revealed. Moreover, three separate one-way ANOVAs revealed that students in the text-only and 2 s-before condition had similar premodels both with regard to the global spatial structure in the corresponding direction, $F < 1$, with regard to the global spatial structure in the non-corresponding direction, $F < 1$, and with regard to the local spatial structure, $F(1, 37) = 1.62, p = .21, \eta_p^2 = .04$. Moreover, 34 out of 39 participants scored 0 for premodels regarding the global spatial structure, suggesting that participants did not represent the pulley system in a particular diagonal direction.

4.2.2. Learning outcomes

One-way ANOVAs with experimental condition as factor (text-only vs. 2 s-before) were conducted to analyze the learning outcomes with respect to postmodels, verbal recall, mental animation, and comprehension of the principle underlying pulley systems. There were no significant differences in learning outcomes between students in the text-only and in the 2 s-before condition as revealed by results for postmodels regarding the global spatial structure in the corresponding direction, $F(1, 37) = 2.78, p = .10, \eta_p^2 = .07$, postmodels regarding the global spatial structure in the non-corresponding direction, $F < 1$, postmodels regarding the local spatial structure, $F < 1$, verbal recall, $F < 1$, mental animation, $F(1, 37) = 1.05, p = .31, \eta_p^2 = .03$, and comprehension of the principle, $F < 1$.

4.2.3. Eye movements

We investigated students’ processing of text by means of analyzing the dwell times spent on reading both the structure and the function section of the text. Therefore, we conducted a $2 \times 2$ ANOVA with condition (text-only vs. 2 s-before) as between-subjects factor, text section (structure vs. function section) as within-subjects factor and dwell time as dependent variable. The ANOVA revealed significant main effects of text section, $F(1, 37) = 120.33, p < .001, \eta_p^2 = .77$, and condition, $F(1, 37) = 4.60, p = .04, \eta_p^2 = .11$. Most importantly, the interaction was significant, $F(1, 37) = 9.68, p = .004, \eta_p^2 = .21$ (see Fig. 5). Post hoc tests revealed that students in the 2 s-before condition attended the structure section for a shorter time period than students in the text-only condition ($p = .01$), whereas they did not differ in their dwell times for the function section ($p = .53$).

4.3. Discussion

In line with Hypothesis 4, students in the 2 s-before condition had shorter reading times on the text section describing the system’s spatial structure than students in the text-only condition, suggesting that processing of text was facilitated by information about the system’s global spatial structure extracted from brief initial picture inspection. However, students in both conditions had similar comprehension levels suggesting that beneficial effects on an outcome-level from briefly inspecting the picture did not apply if text was more specific with regard to the system’s global spatial structure.

Table 5

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Required information</th>
<th>Text-only</th>
<th>2 s-Before</th>
<th>$M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial abilities (min. $= 0$, max. $= 10$)</td>
<td>--</td>
<td>5.60 (1.34)</td>
<td>6.21 (2.46)</td>
<td>5.90 (2.88)</td>
</tr>
<tr>
<td>Premodel (min. $= 0$, max. $= 3$)</td>
<td>Global spatial structure</td>
<td>0.05 (0.22)</td>
<td>0.05 (0.23)</td>
<td>0.05 (0.22)</td>
</tr>
<tr>
<td>Premodel (min. $= 0$, max. $= 1$)</td>
<td>Local spatial structure</td>
<td>0.25 (0.64)</td>
<td>0.05 (0.23)</td>
<td>0.15 (0.49)</td>
</tr>
<tr>
<td>Postmodel (min. $= 0$, max. $= 3$)</td>
<td>Global spatial structure</td>
<td>1.85 (1.39)</td>
<td>2.53 (1.12)</td>
<td>2.18 (1.30)</td>
</tr>
<tr>
<td>Postmodel (min. $= 0$, max. $= 1$)</td>
<td>Local spatial structure</td>
<td>5.85 (4.26)</td>
<td>4.63 (4.51)</td>
<td>5.26 (4.37)</td>
</tr>
<tr>
<td>Verbal recall (min. $= 0$, max. $= 6$)</td>
<td>Global spatial structure</td>
<td>8.45 (3.19)</td>
<td>8.11 (3.07)</td>
<td>8.28 (3.10)</td>
</tr>
<tr>
<td>Verbal recall (min. $= 0$, max. $= 1$)</td>
<td>Local spatial structure</td>
<td>6.14 (0.42)</td>
<td>6.24 (0.42)</td>
<td>6.00 (0.39)</td>
</tr>
<tr>
<td>Mental animation (min. $= 0$, max. $= 10$)</td>
<td>Global spatial structure</td>
<td>5.45 (2.06)</td>
<td>5.47 (2.29)</td>
<td>5.46 (2.15)</td>
</tr>
<tr>
<td>Mental animation (min. $= 0$, max. $= 9$)</td>
<td>Local spatial structure</td>
<td>5.75 (2.45)</td>
<td>4.95 (2.44)</td>
<td>5.36 (2.44)</td>
</tr>
<tr>
<td>Dwell time on structure section</td>
<td>Global + local spatial structure</td>
<td>149.01 (72.41)</td>
<td>99.51 (35.37)</td>
<td>124.90 (61.99)</td>
</tr>
<tr>
<td>Dwell time on function section</td>
<td>Functions</td>
<td>51.19 (30.04)</td>
<td>44.90 (31.51)</td>
<td>48.13 (30.35)</td>
</tr>
</tbody>
</table>
supporting Hypothesis 4, briefly inspecting the picture sped up processing of text about the system's spatial structure compared to processing just text in Experiment 3. This is in line with results obtained in the study of Eitel et al. (2013). One may explain these results by assuming that part of the mental model (regarding the system's global structure) has already been constructed based on brief picture inspection. Thus, instead of having to construct a mental model from scratch when processing text, this part of the mental model did not need to be constructed anymore; rather, it only needed to be reactivated (cf. Glenberg & Langston, 1992). Results regarding Hypothesis 2 support the assumption that the system's global spatial structure, and thus the global structure of the mental model, was reactivated while processing text. Constructing a mental model from text about the spatial structure of the system was thus facilitated, being reflected in shorter reading times on the respective section in the text (cf. Just & Carpenter, 1980; Kaakinen et al., 2003; Rayner et al., 2006).

Taken together, the present findings revealed that processing of a picture — even for a short time only — affected subsequent learning from text, thereby suggesting interplay between text and pictures at an early stage of processing. However, it is important to note that for the sake of internal validity the same, frequently used instructional material (i.e., text and picture of pulley system; cf. Boucheix & Schneider, 2009; Hegarty, 1992; Hegarty & Just, 1993) was used for each of the three experiments reported in the present paper. Thus, it remains an open question whether the findings reported in the present paper will generalize to different instructional materials. For instance, one may ask what happens if more complex learning materials are used. In more complex depictions of causal systems, it may be possible that the system's global spatial structure is not the most salient information. Rather, even irrelevant details may be the most salient information in a complex picture (cf. Lowe, 1999). Thus, whereas in the present experiments students extracted predominantly information concerning the system's global spatial structure from brief initial picture inspection that was relevant to understanding the system's functions, this may not be the case in more complex pictures. Accordingly, using such a complex picture, results from a study of Lowe (1999) revealed that students extracted the perceptually salient rather than the thematically relevant information, which had detrimental effects.
on the learning success. That is the case because salient but thematically irrelevant features in a picture may guide attention away from parts of the picture that are important to learning, possibly corrupting beneficial effects of pictures on comprehension. Accordingly, effects of better comprehension via a brief initial picture inspection — as found in Experiment 2 of the present research and in the study of Eitel et al. (2013) — may be moderated by the degree to which the relevant (global) features are the most salient ones in the picture.

Moreover, even if the (global) spatial structure could be extracted from briefly inspecting a complex causal system picture, it remains to be investigated whether its spatial structure can also be reactivated and maintained in working memory while processing subsequent text. According to Kulhavy et al. (1993), a spatial representation of a coherent picture can be held as a single unit in working memory, thus requiring only little resources. However, if this spatial representation was complex, it might require several units in working memory so that keeping it active might actually exceed working memory capacity during subsequent processing of text. As a result, text-picture integration might fail having detrimental effects on performance (e.g., Mayer, 2009). In line with this reasoning, prior research has shown that when presenting information from text and picture in a sequential manner, the learning success is moderated by the complexity (i.e., element interactivity) of the instructional materials (cf. Ginn, 2006), with more complex instructional materials yielding more detrimental effects in a sequential compared to a simultaneous presentation format. As a consequence, further studies are needed to establish under which conditions, (briefly) inspecting the picture before text yields beneficial effects on learning.

In this regard, it is important to mention that with the present studies we did not aim to find the optimal way to present text and pictures with regard to learning success. Thus, we would not recommend presenting the picture only for a short time before text. Rather, we presented the picture and text in this specific manner to study in isolation how extracting global spatial information from the picture affects comprehension and processing of text. Thus, by no means did the present research aim to challenge the well-established multimedia effect (cf. Mayer, 2009). Rather, the present research aimed to shed more light on multimedia effects when the picture is presented (for a short time) prior to the text.

Nonetheless, initially inspecting the picture instead of initiating the learning process by first reading text has been found to be a learning behavior that fosters comprehension of a multimedia instruction about mitosis and meiosis (Stalbovs, Eitel, & Scheiter, 2013). In addition, in a study of Scheiter and Eitel (2010), introducing signals to a multimedia instruction about the circulatory system led to more and earlier inspection of the picture that, in turn, led to better performance on a subsequent knowledge test. Thus, one may tentatively conclude that even shortly inspecting the picture allows for extracting its global spatial structure which, in turn, may facilitate mental model construction. In line with ITPC (Schnottz & Bannert, 2003), in the present paper it was argued that briefly inspecting the picture allowed for a mapping of the picture’s global spatial structure to the global structure of the respective mental model (analogue structure mapping). Subsequent processing of text was then facilitated, because part of the mental model construction had already been done based on brief picture inspection and subsequent steps of mental model construction from text (via propositional representation) were no longer necessary. Thus, the present results were explained by assuming interplay between text and picture processing on the mental model level as it is assumed by CTML and ITPC (cf. Mayer, 2009; Schnottz & Bannert, 2003).

However, while we argue that even briefly inspecting the picture might provide learners with a preliminary structure of the mental model to facilitate subsequent mental model construction from text, we think it is unlikely that the process of mental model construction has already been completed when inspecting the picture for 600 ms or 2 s. Rather, the time might have been too short to construct a complete and coherent mental representation in the sense of generating a runnable mental model from the picture (Boucheix & Schneider, 2009; Mayer & Gallini, 1990). Thus, we would argue that albeit briefly inspecting the picture provided learners with the preliminary structure of the respective mental model, a coherent mental model that reflects comprehension was constructed only after subsequent processing of the corresponding text. In terms of the ITPC, a coherent mental model would result only if the preliminary mental model from the picture would be elaborated by additional model construction that would be done based on the propositional representation from the text.

Another way of explaining the present results would be to assume interplay between text and picture processing prior to mental model construction. Such a process is indicated in the ITPC diagram by an arrow that points from the surface representation of the picture (i.e., visual perception) to the propositional representation from the text (cf. Schnottz, 2002, p. 109). Thus, it may be that the visual perception of the picture activated propositions that contributed to the propositional representation constructed from text — even before a mental model was constructed from the picture by means of analogical structure mapping. The picture therefore contributed to the mental model by first contributing to the propositional representation that, in turn, provided the basis for mental model construction. In a similar vein, according to CTML interplay might be possible between the two surface representations of text and pictures, namely between words and sounds that were selected from the instruction prior to being organized into separate mental models (Mayer, 2009, p. 61).

Thus, CTML and ITPC allow for several theoretical explanations of the findings obtained within the present studies, and further research is needed that tries to distinguish between these different explanations. Nonetheless, results from the present research specify one instance of how interplay between text and picture processing can happen at an early processing stage.

5.2. Limitations and future studies

In Experiments 2 and 3 of the present research, students were asked to draw a pulley system prior to receiving the instruction. This method was introduced solely to rule out that students’ prior conceptions of pulley systems matched the structure of the to-be-learned system. However, one might argue that this demanding activity prior to the instruction had a significant influence on students’ subsequent learning behavior. For instance, students’ prior conceptions of pulley systems were made explicit by the initial drawing task so that they could have compared each subsequent
piece of information concerning the pulley system to the prior conception. When a pulley system picture was presented in addition to text about the pulley system, the prior conception could be compared more directly to the new information than if just text was presented. Thus, an alternative explanation of the present results might be that the picture provided a more direct feedback to students’ prior conceptions. In consequence, it is yet an open question whether the learning process would be different if no drawing task was given prior to the instruction, which has to be subject to further research. However, in three experiments reported in the present paper as well as in the study of Eitel et al. (2013), results revealed that students’ prior conceptions of pulley systems were not elaborate (i.e., most students received 0 points). Thus, we think it is unlikely that the present results are due to a more direct feedback that the picture gave regarding students’ prior conceptions.

By applying the blank screen paradigm (cf. Altmann, 2004) to learning with text and a picture about pulley systems, in the present research we found evidence that spatial information extracted from previous picture inspection is reactivated when processing text; that is, subsequent text about the pulley system acted as a cue to reactivate previously extracted information concerning the system’s spatial structure. While we found first evidence for this cued reactivation process, with the present materials we were not able to determine the exact time-course of this process. For instance, it remains to be investigated when exactly in the process of learning from text the spatial information extracted from the picture is reactivated, and whether this information is kept in working memory or reactivated several times during the learning process. In line with Kulhavy et al. (1993), one might argue that a holistic image from the initially inspected picture (e.g., the three pulleys in their diagonal orientation) is reactivated and kept in working memory throughout the learning process. In contrast, it is possible that the picture is reactivated piecemeal when processing subsequent text (cf. Hegarty 1992). For instance, as soon as the text refers to the “upper pulley”, learners might reactivate their mental image of the upper pulley, which would be accompanied by eye movements on the blank screen that are made towards the respective location in which the upper pulley had been presented in the picture. To investigate reactivation processes on such a fine-grained level, however, one would need to present study materials in a very artificial way, for instance, by including a pause after each mentioning of the word “upper pulley” in the text to be able to unambiguously relate the eye movements to the respective term in the text.

In the present studies, by contrast, materials were presented at a normal reading speed to have a learning situation that is as realistic as possible. Thus, in the present studies we tried to find the best compromise between studying effects on a process-level and presenting materials as realistic as possible. Results provide tentative evidence for a scaffolding process, meaning that global spatial information, which is extracted even from shortly inspecting the picture, is reactivated and integrated with text to have beneficial effects on comprehension and processing of text.

Acknowledgments

This research is funded by the Pact for Research and Innovation of the Competition Fund of the Leibniz Gemeinschaft. Thanks to Elisabeth Arzberger for data collection help. Thanks to the eye tracking group within the Humanities Lab of Lund University, to Dr. Lucia Mason, and to the three anonymous reviewers for valuable comments regarding drafts of this article.

Appendix A

Test of pictorial relations

Eleven relations were scored in the drawings (adapted from Hegarty, 1992):

- The upper pulley is fixed to the ceiling
- The middle pulley is free to move up and down
- The crate is suspended from the lower pulley
- The upper rope has a free ending
- The upper rope lies over the upper pulley
- The upper rope goes from the upper pulley to the middle pulley
- The upper rope lies under the middle pulley
- The upper rope is attached to the ceiling
- The lower rope is attached to the middle pulley
- The lower rope lies under the lower pulley
- The lower rope is fixed to the ceiling

Appendix B

Text about structure and functions of the pulley system in Experiment 2

The pulley system

The pulley system consists of three pulleys, two ropes and one weight. The upper pulley is attached to the ceiling. Below the upper pulley is the middle pulley that is free to move up and down, and is therefore called free pulley. The upper rope is attached to the ceiling at one end, goes under the middle pulley and over the upper pulley, and is free at the other end. The lower pulley is free to move up and down, and is therefore called free pulley as well. The lower rope is attached to the ceiling at one end. It goes under the lower pulley and is attached to the middle pulley at the other end. The crate is suspended from the lower pulley. When the free end of the upper rope is pulled, the rope moves over the upper pulley and under the middle pulley and pulls up the middle pulley. This causes the lower rope to move under the lower pulley and to pull up both the lower pulley and the crate. For these types of pulley systems, each free pulley that is added to the system splits the force with which the weight has to be lifted in half. Each free pulley that is added to the system, however, also doubles the length of rope to be pulled.

References


